Attention-driven Multi-sensor Selection

Braun, Stefan ; Neil, Daniel ; Anumula, Jithendar ; Ceolini, Enea ; Liu, Shih-Chii

Abstract: Recent encoder-decoder models for sequence-to-sequence mapping show that integrating both temporal and spatial attention mechanisms into neural networks considerably improve network performance. The use of attention for sensor selection in multi-sensor setups and the benefit of such an attention mechanism is less studied. This work reports on a sensor transformation attention network (STAN) that embeds a sensory attention mechanism to dynamically weigh and combine individual input sensors based on their task-relevant information. We demonstrate the correlation of the attentional signal to changing noise levels of each sensor on the audio-visual GRID dataset and synthetic noise; and on CHiME-4, a multi-microphone real-world noisy dataset. In addition, we demonstrate that the STAN model is able to deal with sensor removal and addition without retraining, and is invariant to channel order. Compared to a two-sensor model that weighs both sensors equally, the equivalent STAN model has a relative parameter increase of only 0.09%, but reduces the relative character error rate (CER) by up to 19.1% on the CHiME-4 dataset. The attentional signal helps to identify a lower SNR sensor with up to 94.2% accuracy.

DOI: https://doi.org/10.1109/ijcnn.2019.8852396

Attention-driven Multi-sensor Selection

Stefan Braun, Daniel Neil, Jithendar Anumula, Enea Ceolini, Shih-Chii Liu
Institute of Neuroinformatics
University of Zurich and ETH Zurich
Zurich, Switzerland
brauns@ethz.ch, daniel.l.neil@gmail.com, anumula@ini.uzh.ch, enea.ceolini@ini.uzh.ch, shih@ini.ethz.ch

***Abstract*—Recent encoder-decoder models for sequence-to-sequence mapping show that integrating both temporal and spatial attention mechanisms into neural networks considerably improve network performance. The use of attention for sensor selection in multi-sensor setups and the benefit of such an attention mechanism is less studied. This work reports on a sensor transformation attention network (STAN) that embeds a sensory attention mechanism to dynamically weigh and combine individual input sensors based on their task-relevant information. We demonstrate the correlation of the attentional signal to changing noise levels of each sensor on the audio-visual GRID dataset and synthetic noise; and on CHiME-4, a multi-microphone real-world noisy dataset. In addition, we demonstrate that the STAN model is able to deal with sensor removal and addition without retraining, and is invariant to channel order. Compared to a two-sensor model that weighs both sensors equally, the equivalent STAN model has a relative parameter increase of only 0.09%, but reduces the relative character error rate (CER) by up to 19.1% on the CHiME-4 dataset. The attentional signal helps to identify a lower SNR sensor with up to 94.2% accuracy.**

***Index Terms*—multi-sensor input, attention mechanism, sensor fusion, end-to-end speech recognition**

I. Introduction

Modern robotic systems handle complex interactions with the real world. For example, recent research shows how drones can navigate autonomously in natural environments [1], how rescue robots can operate in degraded environments [2], [3], and how robots can interact with humans via physical contact [4] or speech [5]. Such interactions require an informative perception of the environment for appropriate action. Even though single-sensor setups have been proposed for tasks like autonomous driving [6], [7], it is advantageous to equip robots with multiple sensors of the same or distinct modalities to increase the performance, robustness and fault tolerance of the system [8].

With deep learning methods, robotic interaction can be modelled as a domain-agnostic mapping from multiple input sensors to a task-specific interaction with the environment. End-to-end models handle the input-output mapping with a single neural network that is optimized on large amounts of training data. Compared to traditional pipelined approaches with separately optimized stages, end-to-end models represent a significant simplification in model complexity and optimization process. However, non-trivial architectural decisions remain such as *where* and *how* to combine the information from the multiple inputs in a single network. The combination process is generally referred to as *sensor fusion*, and neural networks offer various fusion strategies such as fusion at different network layers (where to fuse) and different fusion operations such as concatenation, summation and convolution (how to fuse) [9].

In this work, we investigate a recently proposed fusion operation based on *sensory attention* [10], [11]. The attention-based fusion operation first weighs and then sums multiple input sensors into a single representation. The sensor-specific attention weights are computed by neural networks that are integrated into the end-to-end model such that a single training process is sufficient. This attention mechanism allows the model to dynamically tune its attention towards more task-informative sensors.

We evaluate the use of this sensory attention mechanism when embedded in a sensor transformation attention network (STAN). STANs support multi-sensor inputs of the same or different modalities in a single end-to-end framework (Section II). We test STANs in noisy conditions using two different datasets, the multi-modal GRID dataset with added synthetic noise (Section III), and the multi-microphone CHiME-4 dataset with real-world noise (Section IV). We show that the attention weights are highly interpretable and that our STAN model is able to deal with new sensor configurations without any re-training, including sensor addition, removal and reversal of the sensor order.

A. Related work

Attention mechanisms have contributed to improved results from deep neural networks (DNNs) in various domains such as image captioning [12], video description [13], speech recognition [14], [15], and machine translation [16], [17]. These attention mechanisms enable DNNs to sequentially focus on different subsets of an input source [18]. Typical examples include *spatial attention*, e.g. aligning output words to regions within an image during caption generation [12] or *temporal attention*, e.g. aligning output words to distinct feature frames of the input sequence in speech recognition [14].

Less work is reported on the use of *sensory attention*. Two recent studies evaluated sensory attention in the context of end-to-end models for multi-channel automatic speech recognition (ASR) [10], [19]. In both studies, inputs from multiple channels are combined into a single representation that is used for classification. The sensory attention mechanism is either used to predict the reference microphone for a neural

beamformer [19] or as a fusion operation [10]. Similar to [10], we use sensory attention as a fusion operation, but follow a different design strategy. Our proposed design improves with invariance to channel order and the design is simplified by using long short-term memory (LSTM) and dense units instead of a custom-designed neural network cell. For multi-modal inputs, the only study we are aware of that uses sensory attention was carried out by [11]. Their work used a different dataset (YouTube2Text) and did not quantify the correlation of the attentional signal with the noise level.

The use of sensory attention is notably absent from other studies on multi-modal end-to-end models. A recent study evaluated the impact of different fusion operations such as summation, concatenation, convolution, max-pooling and a bilinear operator for a video classification task [9]. Other recent studies that evaluated gesture recognition [20], [21] or robot control (autonomous car racing [22], search and pick [23]) also used other fusion operations in their end-to-end models.

II. Sensor Transformation Attention Network

The STAN architecture depicted in Figure 1 includes the following building blocks: (1) input sensors, (2) sensor transformation functions and (3) a sensory attention mechanism. The attention mechanism combines multiple input sensors into a single, merged representation by first weighting and then summing transformed feature frames from individual sensors.

We assume a multi-sensor setup with $i = 1, ..., N$ sensors. All sensors record time series that are binned into $k = 1, ..., K$ frames, such that every sensor s^i provides a D_f-dimensional feature vector $f_k^i \in \mathbb{R}^{D_f}$ for each frame k. The merged representation $m \in \mathbb{R}^{D_t}$ is generated by the steps described in Eqs. 1 to 4:

$$t_k^i = T^i(f_{1..k}^i) \tag{1}$$

$$z_k^i = Z^i(t_{1..k}^i) \tag{2}$$

$$\alpha_k^i = \frac{\exp(z_k^i)}{\sum_{j=1}^{N} \exp(z_k^j)} \tag{3}$$

$$m_k = \sum_{i=1}^{N} \alpha_k^i \cdot t_k^i \tag{4}$$

The transformation function T^i converts the feature vectors f_k^i to transformed feature vectors $t_k^i \in \mathbb{R}^{D_t}$ (Eq. 1). If no transformation is desired, then T^i is the identity function $f_k^i = t_k^i$. The attention scoring function Z^i produces scalar attention scores $z_k^i \in \mathbb{R}^1$ based on the transformed features of sensor i (Eq. 2). The attention weights $\alpha_k^i \in \mathbb{R}^1$ are computed by performing a softmax operation over all attention scores $z_k^i \in \{z_k^1, ..., z_k^N\}$ (Eq. 3), and therefore $\sum_{i=1}^{N} \alpha_k^i = 1$. Each transformed feature vector t_k^i is then scaled by the corresponding attention weight α_k^i and merged through a summation operation (Eq. 4). The resulting - transformed, scaled and merged - feature vectors m_k are then presented to the classifier.

The sensory attention model implements the scoring functions Z^i which can be modelled using neural networks. For our experiments, we implemented Z^i using 20 LSTM units [1] [24] followed by one dense unit (weight W, bias b) with a SELU non-linearity [2] [25] (Eq. 5). LSTM units are a convenient choice because past history is automatically considered.

$$Z^i(t_{1..k}^i) = \text{SELU}(W \cdot \text{LSTM}(t_{1..k}^i) + b) \tag{5}$$

Our sensory attention mechanism has the following useful properties: First, it is a *soft* attention mechanism, therefore it is differentiable and trainable with back-propagation. Second, at each frame k, the attention weights α_k^i sum up to 1 across all sensors, indicating the contribution of single sensors to the combined representation of a frame. Third, because the attention scores (z_k^i) and weights (α_k^i) are computed on every frame, their values reflect the dynamic per frame adjustment for temporal changes in signal quality due to noise, sensor failure or other sensor corruptions. Finally, because the attention scoring function of each sensor is independently evaluated, existing sensors may be removed or new sensors may be added after training. Note that the scoring function Z^i of each sensor may be identical when their parameters are shared ($\theta_{Z^1} = \theta_{Z^2} = ... = \theta_{Z^N}$). In the shared case, the attention mechanism would then be invariant to sensor re-ordering because the same scoring function is used for all sensors.

The same arbitrary choice of functions can be made for the transformation functions T^i as in the scoring functions Z^i. In this work, we used dense units or identity functions. However, other network types such as convolutional neural networks (CNNs) [26] might work just as well or even better.

III. Experiments with multi-modal input

To evaluate the impact of non-stationary noise on the attention mechanism in a multi-modal setup, we developed experiments on the audio-visual GRID dataset. These experiments with controlled noise levels allowed us to establish the ground truth in measuring the correlation between the attention weights and the signal-to-noise ratio (SNR) of the input sensors.

A. Dataset

The GRID dataset [27] provides audio and video (facial) recordings of 1000 sentences each spoken by 34 speakers. The recording setup consisted of a single camera and microphone. Each sentence contains 6 vocabulary units out of a word vocabulary of 51 classes (commands, colors, prepositions, adverbs, letters and digits). The samples were shuffled and split by 80/10/10% into training, validation and test sets. The

[1] We found ranges from 20 to 200 LSTM units perform equally well in our experiments.

[2] The choice of the SELU activation function is arbitrary. We tried different variants (ReLU, LeakyReLU, SELU) and they all worked equally well. The normalization effect from SELU is not required.

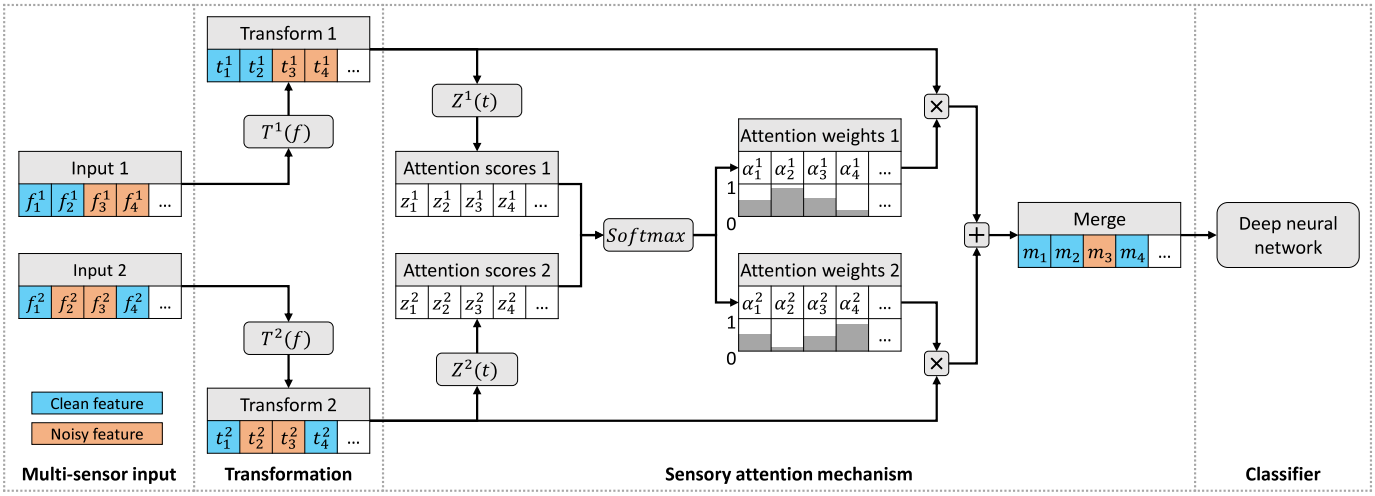

Fig. 1: STAN architecture for a setup with two input sensors. The input feature vectors f_k^i are transformed and then weighted and summed to generate the merged representation m_k that is used for classification. The sensory attention mechanism dynamically adapts its attention weights to create a cleaner merged representation.

raw audio was converted to 123 dimensional filterbank features (40 Mel-spaced filterbanks, energy coefficient, 1st and 2nd order delta features, 50ms frames, 25ms frame shift). The video recordings were processed to extract 17x8 pixel-sized mouth crops by using the Dlib face detector and a pre-trained model of the 68 facial landmark annotator [28], [29]. The pre-processed features of both modalities were presented to the networks with the same frame rate and no further temporal alignment efforts were applied. Both feature types were zero-mean and unit-variance normalized on a per-sample basis. The task in our GRID experiments is ASR: sequences of input features are transcribed to sequences of words. The word error rate (WER) was used as the performance metric.

B. Models

Five models are evaluated: the uni-modal, single-sensor (1) `AUDIO` and (2) `VIDEO` models and the multi-modal, two-sensor (3) `CONCAT`, (4) `AVG` and (5) `STAN` models. All models convert each of their input sensors with individual transformations T^i (Eq. 1) that are implemented with 50 dense units followed by a SELU non-linearity [25], therefore $t_k^i \in \mathbb{R}^{50}$. The multi-modal networks combine the transformed features t_k^i from each sensor by concatenation (`CONCAT`), averaging (`AVG`) or with the sensory attention mechanism (`STAN`). The `STAN` scoring functions $\{Z^1, Z^2\}$ are both implemented with 20 LSTM units followed by a single dense unit with a SELU non-linearity (Eq. 5). Because the sensors provide distinct feature modalities, the parameters of the scoring functions are not shared, i.e. $\theta_{Z^1} \neq \theta_{Z^2}$. All models use the same classifier architecture consisting of 2 layers of bidirectional LSTM cells with 200 units per direction followed by a 52-dimensional output projection, but are individually optimized. The models are trained in an end-to-end fashion with the Connectionist Temporal Classification (CTC) objective [30] and the ADAM optimizer [31] for 150 epochs. The model achieving the lowest WER on the validation set is used for evaluation. Note that there is no modality-specific pre-training, and instead all models are trained from scratch with randomly initialized weights.

C. Noise models

The clean data provided by either modality provides reasonable information to solve the ASR task on GRID, and a `STAN` model with clean multi-modal input did not show much attentional switching in consequence. To encourage attention switching, we add synthetic noise to each feature frame f_k^i. The noise is sampled from a zero-centered uniform distribution with standard deviation $\sigma(k)$, that we refer to as the noise level. We use three different noise models: (1) random walk noise, (2) cross noise and (3) hi-lo noise, with examples shown in Figure 2.

The *random walk noise* model adds noise with a time-varying noise level, $\sigma(k)$, to each sensor and is used for both training and testing. The random walk process $q(k)$ is drawn separately for each sensor, and $q(k)$ is then normalized to the range $[0, 1]$ and scaled by the maximum noise level σ_{max} as shown in Eqs. 6 and 7. As a result, each sensor has its own distinct noise level at each time step.

$$q(k) = \sum_{i=1}^{k} n_i, \text{ with } n_i \sim \mathcal{N}(0, 1) \tag{6}$$

$$\sigma(k) = \sigma_{max} \cdot \frac{q(k) - \min\{q_1, ..., q_K\}}{\max\{q_1, ..., q_K\} - \min\{q_1, ..., q_K\}} \tag{7}$$

The resulting random walk process yields an average noise level of $E(\sigma(k)) = \sigma_{max}/2$. The training noise level $\sigma_{max} = 8$ is chosen such that the average WER is close to 16% when using a single modality. This allowed the multi-sensor models to have a good chance of improvement by considering the other modality, while at the same time a single modality still

provided reasonable WER with 5 correct words out of 6. We noticed that this noise level was necessary to encourage attentional switching for STANs.

The *cross* and *hi-lo* noise variants are only used during testing so we can evaluate the generalization of the attention mechanism. The cross noise applies a linearly increasing noise level $\sigma(k) = \sigma_{max} \cdot {}^{k}/_{K}$ to one sensor, and a linearly decreasing noise level $\sigma(k) = \sigma_{max} \cdot (1 - {}^{k}/_{K})$ to the other sensor. The hi-lo noise applies a constant noise level $\sigma(k) = \sigma_{max}$ to one sensor, and the noise level $\sigma(k) = 0$ to the other sensor. The sensor that sees the increasing (cross noise) or maximum (hi-lo) noise level is alternated for every test sample. In consequence, either sensor sees the increasing or maximum noise level for 50% of all test samples. All models are tested with the same alternation pattern.

D. Attention metrics

Because the noise level of both sensors is known at any time, the interpretability of the attention weights can be quantified. One proposed metric, attention correlation (ATTCORR), measures the correlation between the noise level and the attention weights for a specific sensor. We use the notation sensor index i (audio=1, video=2), noise level σ^i and attention weights α^i to define the ATTCORR as a correlation coefficient with values between -1 and 1:

$$\text{ATTCORR}^i = corr\Big(\big(1 - 2 \cdot \frac{\sigma^i}{\sum_{j=1}^{2} \sigma^j}\big), \big(2 \cdot \alpha^i - 1\big)\Big) \quad (8)$$

A value of 1 corresponds to perfect correlation, and 0 corresponds to chance level. Another quantitative metric is the attention accuracy (ATTACC) which measures the accuracy of identifying the higher (or lower) SNR sensor by their attention weights:

$$\text{ATTACC}\,[\%] = 100 \cdot \frac{f_{correct}}{f_{total}} \quad (9)$$

where $f_{correct}$ is the number of frames with correct SNR sensor identification and f_{total} is the total number of frames of the evaluation set. Sensors are considered as correctly identified on a frame when the lower SNR sensor is attributed a lower attention weight. An ATTACC value of 100% corresponds to perfect identification, and a value of 50% corresponds to chance level.

E. Results

The models are trained in noisy conditions with added random walk noise ($\sigma_{max} = 8$), and evaluated in clean and noisy conditions with random walk, cross or hi-lo noise added. Table I reports the WER for all five models; and the ATTACC and ATTCORR scores for the `STAN` model.

The multi-modal networks perform significantly better than the uni-modal networks in both clean and noisy conditions. The `STAN` and `AVG` models are mostly on par and achieve the lowest WER, except for the hi-lo noise where `STAN` shows a relative WER improvement of up to 36.1% ($\sigma_{max} = 8$) over `AVG`. Across all noise types and all noise levels, `STAN` shows a relative WER improvement of 9.1% to 36.9% over `CONCAT`, 56.8% to 77.7% over `AUDIO` and 64.1% to 82.8% over `VIDEO`.

The `STAN` model computes highly interpretable attention weights as shown in Figure 2. Both ATTACC and ATTCORR metrics start at chance level for $\sigma_{max} = 1e-5$ and increase to scores between 75.3% to 99.8% (ATTACC) and 0.68 to 0.83 (ATTCORR) for $\sigma_{max} = 8$. Failures in correct prediction of the lower SNR sensors mainly arise when both sensors have similar noise levels, or noise levels change too rapidly. These cases are mostly seen for the random walk noise. Even though the model is trained only on random walk noise, the attention mechanism generalizes well to the cross and hi-lo noise types.

IV. Multi-Channel Speech Recognition with Natural Noise

We further evaluate STANs on the multi-channel ASR CHiME-4 dataset which includes *real-world* noise. The reported experiments include a comparison of STANs against concatenation, averaging and beamforming models. Without any re-training, we further evaluate the robustness of our models with respect to reversed channel orders, channel addition and channel removal.

A. Dataset

The CHiME-4 dataset [32] for ASR provides real and simulated noisy speech data from a tablet device with 6 microphones. The data was recorded in four noisy environments: public transport, a cafe, a street junction and pedestrian area. The real data was recorded with the tablet device, and the simulated data was generated by mixing clean speech utterances from the WSJ0 [33] dataset with environment background recordings. The tablet device is equipped with 5 microphones facing the speaker and 1 microphone facing away from the speaker (backward channel #2, the noisiest of all). We use both real data (`tr05_real`, 1600 samples) and simulated data (`tr05_simu`, 7138 samples) for training, and evaluate on the real noisy data subsets (`et05_real`, 1320 samples and `dt05_real`, 1640 samples). The samples were pre-processed into 123-dimensional filterbank features (40 Mel-spaced filterbanks, energy coefficent, 1st and 2nd order delta features, 25ms frames, 10ms frame shift) and normalized to zero-mean and unit variance per sample. The output labels consist of 59 alphabet units (characters, digits etc.) that are obtained with the EESEN pre-processing routines [34]. The task in our CHiME-4 experiments is ASR: sequences of input features are transcribed to sequences of characters. The character error rate (CER) is used as the performance metric.

B. Models

In total, five different models are evaluated: `CONCAT-2CH`, `AVG-2CH`, `STAN-2CH`, `STAN-5CH` and `BEAMFORMIT-5CH`. The two-channel models are trained on channels (2,5), with the low SNR backwards channel 2 and the high SNR front channel 5 [32]. The five-channel models are trained on the five front channels (1,3,4,5,6). Each

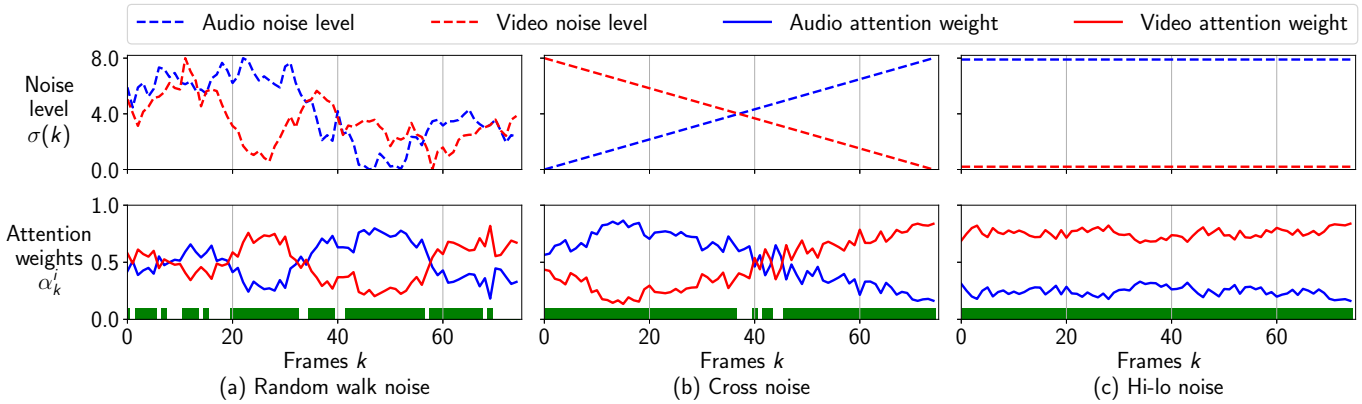

Fig. 2: Attention response on a randomly selected sample from the multi-modal GRID dataset. The top row depicts the noise levels applied to each input, and the bottom row depicts the attention weights computed by STAN. The green bars indicate frames where the relative SNR value of the correct sensor is identified. (a) shows the response to random walk noise resulting in ATTACC of 72%. Note how the attention weights dynamically change, mostly in correlation with the noise level. (b) and (c) show responses to cross and hi-lo noise, with ATTACCs of 92% and 100% respectively.

TABLE I: Results of the GRID experiments, averaged over 10 runs. All values are reported in the format *mean* ± *standard deviation*. The ATTCORR values are not computed for the hi-lo noise because the correlation function is not defined for constant functions. The lowest WER is printed bold. The ATTCORR and ATTACC values are rescaled to the range $[-100, 100]$ in the interest of readability.

Metric	min, max, chance level	Model	Random walk noise, σ_{max} 1e-5	4	8	Cross noise, σ_{max} 1e-5	4	8	Hi-lo noise, σ_{max} 1e-5	4	8
WER [%]	0, 100, 98	AUDIO	2.8 ± 0.1	7.9 ± 0.3	18.5 ± 0.5	2.8 ± 0.1	7.5 ± 0.3	19.1 ± 0.6	2.8 ± 0.1	10.7 ± 0.3	23.8 ± 0.5
		VIDEO	5.2 ± 0.5	9.4 ±0.7	22.3 ± 0.9	5.2 ± 0.5	9.2 ± 0.7	22.0 ± 0.8	5.2 ± 0.5	13.3 ± 0.6	30.8 ± 0.4
		CONCAT	1.1 ± 0.2	2.9 ± 0.6	9.4 ± 1.2	1.1 ± 0.2	2.6 ± 0.5	8.8 ± 1.2	1.1 ± 0.2	3.2 ± 0.6	8.4 ± 1.2
		AVG	**1.0 ± 0.1**	**2.3 ± 0.2**	8.2 ± 0.4	**1.0 ± 0.1**	**2.1 ± 0.2**	7.4 ± 0.4	**1.0 ± 0.1**	2.8 ± 0.2	8.3 ± 0.7
		STAN	**1.0 ± 0.1**	2.4 ± 0.1	**8.0 ± 0.2**	**1.0 ± 0.1**	2.2 ± 0.1	**7.1 ± 0.3**	**1.0 ± 0.1**	**2.6 ± 0.1**	**5.3 ± 0.2**
ATTACC [%]	0, 100, 50	STAN	50.6 ± 0.1	59.7 ± 1.4	75.3 ± 1.1	49.2 ± 0.1	65.1 ± 2.1	87.6 ± 1.5	50.0 ± 0.1	87.3 ± 2.4	99.8 ± 0.1
ATTACC/scale	-100, +100, 0	STAN	1.2 ± 0.1	19.4 ± 2.8	50.6 ± 2.2	-1.6 ± 0.1	30.2 ± 4.2	75.1 ± 3.0	0.0 ± 0.1	74.6 ± 4.8	99.7 ± 0.1
ATTCORR[1]	-100, +100, 0	STAN	1.2 ± 0.1	33.7 ± 2.8	68.0 ± 1.2	-0.5 ± 0.1	53.6 ± 3.2	83.4 ± 1.8	-	-	-
ATTCORR[2]	-100, +100, 0	STAN	1.3 ± 0.1	33.8 ± 2.7	68.2 ± 1.2	-0.5 ± 0.1	53.6 ± 3.2	83.4 ± 1.8	-	-	-

input channel provides pre-processed filterbank features, and because it was not necessary to use further transformations, we choose the identity transformation function T^i ($t_k^i = f_k^i$). The models apply different channel combination strategies: the CONCAT-2CH model concatenates both input channels for classification, and the AVG-2CH averages both input channels before classification. The averaging strategy corresponds to assigning fixed attention weights $\alpha_k^i = 1/2$ to the input frames. The STAN models compute data-dependent attention weights and implement the channel scoring functions Z^i with 20 LSTM units followed by a single dense unit with a SELU non-linearity (Eq. 5), resulting in 11k additional parameters over the AVG model (+0.09% relative). Because the input channels are of the same modality, we apply the same scoring function Z to each channel i, therefore $\theta_{Z^1} = ... = \theta_{Z^N}$. The BEAMFORMIT-5CH model uses a delay-and-sum beamformer [35] which first produces enhanced waveforms in a separate pre-processing stage so it is not considered as an end-to-end model.

All five models use the same classifier architecture based on 5 layers of bidirectional LSTM cells with 350 units per direction followed by a 59-dimensional output projection, but are individually optimized. The models are trained in an end-to-end fashion for 150 epochs with the ADAM optimizer and the CTC objective. The model with the lowest CER on the development set was used for evaluation.

We further include results from related models on multi-channel end-to-end ASR without additional lexicons or language models. The ATTMULTI-E2E [10] combines multiple input channels into a single representation with a sensory attention mechanism based on weighted summation. Their attention mechanism shows two main differences to the one used in our STAN models: (1) it uses a custom designed neural network cell to compute attention scores while we use generic LSTM and dense units and (2) it is not invariant to the re-ordering of input channels, while ours is invariant due to the choice of $\theta_{Z^1} = ... = \theta_{Z^N}$. The MASK_NET(ATT) [19] model uses an attention mechanism that selects the reference microphone for a neural beamformer. In contrast to the ATTMULTI-E2E and STAN models, the

TABLE II: Results for the CHiME-4 multi-channel ASR experiments. The CER [%] is given for the `et05_real` and `dt05_real` subsets. The attention weights for `STAN-2CH` and `STAN-5CH` are averaged over all frames of the `dt05_real` subset. The lowest CER and highest attention weight are printed bold. All models are trained and tested on matched channel configurations, and the `CONCAT`, `AVG` and `STAN-2CH` models are additionally tested on new channel configurations without re-training.

ID	Model	Train channels	Test channels	Parameters	et05 real	dt05 real	STAN attention weights, dt05_real $\bar{\alpha}^1$	$\bar{\alpha}^2$	$\bar{\alpha}^3$	$\bar{\alpha}^4$	$\bar{\alpha}^5$	$\bar{\alpha}^6$
(a)	CONCAT-2CH	2,5	2,5	13.508M	**30.4**	20.4	-	-	-	-	-	-
(a)	AVG-2CH	2,5	2,5	13.154M	36.3	24.6	-	-	-	-	-	-
(a)	STAN-2CH	2,5	2,5	13.165M	**30.4**	**19.9**	-	0.22	-	-	**0.78**	-
(b)	CONCAT-2CH	2,5	5,2	13.508M	57.8	43.1	-	-	-	-	-	-
(b)	AVG-2CH	2,5	5,2	13.154M	36.3	24.6	-	-	-	-	-	-
(b)	STAN-2CH	2,5	5,2	13.165M	**30.4**	**19.9**	-	0.22	-	-	**0.78**	-
(c)	AVG-2CH	2,5	1,3,4,5,6	13.154M	**24.9**	**17.2**	-	-	-	-	-	-
(c)	AVG-2CH	2,5	1,2,3,4,5,6	13.154M	27.0	18.5	-	-	-	-	-	-
(c)	STAN-2CH	2,5	2	13.165M	61.3	47.3	-	**1.00**	-	-	-	-
(c)	STAN-2CH	2,5	5	13.165M	28.8	19.3	-	-	-	-	**1.00**	-
(c)	STAN-2CH	2,5	1,3,4,5,6	13.165M	25.7	17.4	0.19	-	0.18	0.19	**0.23**	0.21
(c)	STAN-2CH	2,5	1,2,3,4,5,6	13.165M	26.4	17.8	0.17	0.07	0.17	0.19	**0.21**	0.19
(d)	STAN-5CH	1,3,4,5,6	1,3,4,5,6	13.165M	26.5	17.7	0.17	-	0.17	0.21	**0.23**	0.22
(d)	BEAMFORMIT-5CH	1,3,4,5,6	1,3,4,5,6	13.154M	**24.2**	**15.9**	-	-	-	-	-	-
(d)	ATTMULTI-E2E [10]	1,3,4,5,6	1,3,4,5,6	∼8M	38.0	26.8	-	-	-	-	-	-
(d)	MASK_NET(ATT) [19]	1,3,4,5,6	1,3,4,5,6	∼18M	26.8	18.2	-	-	-	-	-	-

channels are not combined by a sensory attention mechanism, but rather by a neural beamformer. The neural beamformer is able to exploit spatial information, which is not considered by `ATTMULTI-E2E` and `STAN`. Both `ATTMULTI-E2E` and `MASK_NET(ATT)` use a CTC+Encoder/Decoder hybrid model that is trained with a joint CTC-attention multi-task objective, while the `STAN` model is trained with an encoder (i.e. the acoustic model) and single CTC objective.

C. Results

The evaluation is carried out on the `et05_real` and `dt05_real` subsets and the CER results are reported in Table II. For `STAN-2CH` and `STAN-5CH`, we also report the average attention weight $\left(\bar{\alpha}^i = \frac{1}{K}\sum_{k=1}^{K}\alpha_k^i\right)$ of every input channel obtained on the `dt05_real` subset ($K = 985619$ frames). Note that the way we report the attention weights corresponds to the physical CHiME-4 channels, and does not reflect the input channel order.

1) Two-channel models & matched channel order: The models are trained and tested on the channel order (2,5) (see Table II(a)). The `STAN-2CH` and the `CONCAT-2CH` models perform best and achieve similar error rates. `STAN-2CH` shows a relative CER improvement between 16.3% to 19.1% over `AVG-2CH`. Seemingly, `STAN-2CH` benefits from the automatically learned channel weighting. The average attention weight assigned to channel 5 is 3.5x higher than for the noisy channel 2: $\bar{\alpha}^5 = 3.5 \cdot \bar{\alpha}^2$, and the relation $\alpha_k^5 > \alpha_k^2$ holds true for 94.2% of all $K = 985619$ frames. In other words, by comparing attention weights alone, we can identify the higher SNR channel 5 with 94.2% accuracy.

2) Two-channel models & reversed channel order: The models were originally trained on channel order (2,5) but are then tested on the reversed channel order (5,2) without any re-training (Table II(b)). As expected, the `STAN-2CH` and `AVG-2CH` models show error rates that are identical to the train channel order, as both are invariant to channel order. `CONCAT-2CH` performs worse with the reversed channels and shows a relative CER increase of 90.1% to 111.3% compared to the train channel order.

3) Channel addition and removal: The `STAN-2CH` and `AVG-2CH` models were originally trained on channels (2,5), but are then tested on new channel configurations without any re-training (Table II(c)). Both models allow the re-use of the same classifier because the merged feature dimensionality does not change with the number of input channels. The `CONCAT-2CH` model is not considered here as it does not allow to re-use of the classifier: the concatenated feature dimensionality grows with the number of input channels, but the classifier expects the same input dimensionality as during training.

Interestingly, both `AVG-2CH` and `STAN-2CH` show improved CER scores when tested with the new channel configurations (1,3,4,5,6) and (1,2,3,4,5,6), without any re-training. Compared to the channel configuration (2,5), `STAN-2CH` shows relative CER improvements between 10.6% to 15.5% and `AVG-2CH` shows relative CER improvements between 24.8% to 31.4%. Both models now achieve similar CER, and the previous advantage of `STAN-2CH` in the two-channel tests is reduced. This is expected, as the benefit of dynamic channel weighting should be smaller when the merged representation mainly (1,2,3,4,5,6) or only (1,3,4,5,6) consists of the five front channels with similar signal quality. The average attention weights computed by `STAN-2CH` seem reasonable with the five front channels at equal levels. When available, the backwards facing channel 2 can be identified by its significantly lower average attention weight $\bar{\alpha}^2$. We further illustrate the

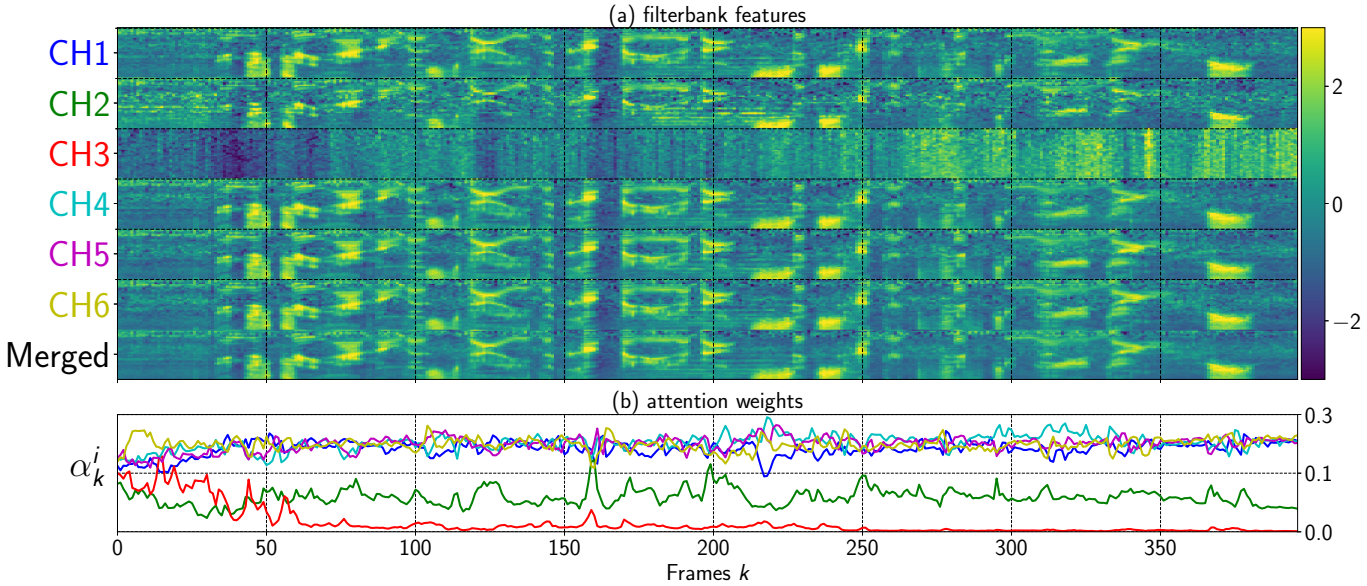

Fig. 3: Operation of STAN−2CH on a sample with channel configuration (1/2/3/4/5/6). (a) Filterbank features for the 6 input channels and the merged representation. (b) Attention weights α_k^i for the 6 input channels. The attention weights show three distinct tiers: the cleanest channels (1,4,5,6) are assigned the highest attention weights that are roughly equal for all 4 channels. The weights of noisy channel 2 lies between those of (1,4,5,6) and the highly corrupted channel 3 (isolated case of microphone failure). The merged representation appears to be hardly corrupted by channels 2 and 3.

operation of the STAN−2CH attention mechanism on a sample using channel configuration (1,2,3,4,5,6) and with channel corruption on channels 2 and 3 in Figure 3.

The single-channel tests are only performed for STAN−2CH, and keeping only channel 5 results in similar CER compared to the default (2,5) configuration. This is no surprise considering that the sensory attention mechanism already favored channel 5. When keeping only channel 2, the CER becomes significantly worse.

4) Baselines: All baseline models are trained and tested on the five front channels (Table II(d)). STAN−5CH shows error rates that are similar to MASK_NET(ATT) [19] and that are significantly lower than ATTMULTI−E2E [10].

The BEAMFORMIT−5CH model achieves the lowest overall error rates with relative CER improvements of 8.7% to 10.2% over STAN−5CH and 5.8% to 8.6% over STAN−2CH. While the beamforming approach shows lower error rates than the sensory attention mechanism, it uses significantly more processing time. In order to generate the enhanced output from the five input channels on a sample of average length 6s, the beamforming algorithm takes 3554ms (CPU) while the attention mechanism of STAN−5CH only takes 195ms (CPU) or 25ms (GPU), i.e. 25x to 142x faster (Skylake Xeon CPU with 4.3GHz, GTX 1080 GPU).

V. Discussion

In this work we presented the STAN end-to-end model that embeds a sensory attention mechanism for sensor fusion in multi-sensor setups. The attention mechanism dynamically decreases attention weights on low SNR sensors, and the STAN model is able to deal with sensor removal and addition without re-training. These properties make the model useful for multi-sensor systems on real-world robotic platforms (e.g. rescue robots [2], [3]) because the attention weights can help to identify failing sensors for replacement or sub-optimal sensors for removal in order to save hardware, computation and energy resources. Furthermore, the attention weights may be informative cues for active perception [36], e.g. helping robots to orientate their head-mounted microphones towards a sound source for improved auditory perception. Also, the sensory attention mechanism may be configured to show invariance towards sensor re-ordering, therefore simplifying the wiring and setup of the sensors on multi-sensor platforms.

The attention model is trainable in an end-to-end fashion and consists of a combination of standard neural network units receiving sensory input only. Therefore this model can be used easily for any new sensor setup, dataset or task other than ASR. Across all the experiments in the paper, the sensory attention mechanism performed on par or better than the concatenation or averaging strategies for sensor combination, with the additional benefit of computing highly interpretable attention weights. Compared to a two-sensor model that weighs both sensors equally, the equivalent STAN model has only a relative parameter increase of 0.09%, but reduces the relative CER by up to 19.1% on the CHiME-4 dataset. Furthermore, the attentional signal of STAN enabled us to identify the lower SNR sensor with up to 94.2% accuracy.

VI. Acknowledgements

This work was partially supported by Samsung Advanced Institute of Technology and the European Union's Horizon 2020 research and innovation program under grant agreement No 644732.

References

[1] A. Giusti, J. Guzzi, D. C. Cireşan, F.-L. He, J. P. Rodríguez, F. Fontana, M. Faessler, C. Forster, J. Schmidhuber, G. Di Caro *et al.*, "A machine learning approach to visual perception of forest trails for mobile robots," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 661–667, 2016.

[2] M. Spenko, S. Buerger, and K. Iagnemma, *The DARPA Robotics Challenge Finals: Humanoid Robots To The Rescue*. Springer, 2018, vol. 121.

[3] Y. Bando, T. Mizumoto, K. Itoyama, K. Nakadai, and H. G. Okuno, "Posture estimation of hose-shaped robot using microphone array localization," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013, pp. 3446–3451.

[4] A. Bajcsy, D. P. Losey, M. K. O'Malley, and A. D. Dragan, "Learning robot objectives from physical human interaction," in *Conference on Robot Learning (CoRL)*, 2017, pp. 217–226.

[5] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5610–5614.

[6] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Ziebam, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[7] P. Drews, G. Williams, B. Goldfain, E. A. Theodorou, and J. M. Rehg, "Aggressive deep driving: Combining convolutional neural networks and model predictive control," in *Conference on Robot Learning (CoRL)*, 2017, pp. 133–142.

[8] H. Cho, Y.-W. Seo, B. V. Kumar, and R. R. Rajkumar, "A multi-sensor fusion system for moving object detection and tracking in urban driving environments," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 1836–1843.

[9] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 1933–1941.

[10] S. Kim and I. Lane, "End-to-end speech recognition with auditory attention for multi-microphone distance speech recognition," in *Interspeech*, 2017, pp. 3867–3871.

[11] C. Hori, T. Hori, T.-Y. Lee, K. Sumi, J. R. Hershey, and T. K. Marks, "Attention-based multimodal fusion for video description," *arXiv preprint arXiv:1701.03126*, 2017.

[12] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 2048–2057.

[13] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4507–4515.

[14] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4945–4949.

[15] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[17] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[18] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1875–1886, 2015.

[19] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, 2017.

[20] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Moddrop: adaptive multi-modal gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, 2016.

[21] F. Li, N. Neverova, C. Wolf, and G. Taylor, "Modout: Learning to fuse modalities via stochastic regularization," *Journal of Computational Vision and Imaging Systems*, vol. 2, no. 1, 2016.

[22] G. Liu, A. Siravuru, S. Prabhakar, M. M. Veloso, and G. Kantor, "Learning end-to-end multimodal sensor policies for autonomous navigation," in *Conference on Robot Learning (CoRL)*, 2017, pp. 249–261.

[23] S. Bohez, T. Verbelen, E. D. Coninck, B. Vankeirsbilck, P. Simoens, and B. Dhoedt, "Sensor fusion for robot control through deep reinforcement learning," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 2365–2370.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017, pp. 972–981.

[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[27] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[28] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: Sentence-level lipreading," *arXiv preprint arXiv:1611.01599*, 2016.

[29] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: database and results," *Image and Vision Computing*, vol. 47, pp. 3 – 18, 2016.

[30] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine learning (ICML)*, 2006, pp. 369–376.

[31] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, 2016.

[33] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete, LDC93S6A," *Linguistic Data Consortium, Philadelphia*, 2007.

[34] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 167–174.

[35] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[36] K. Nakadai, T. Matsui, H. G. Okuno, and H. Kitano, "Active audition system and humanoid exterior design," in *2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2000, pp. 1453–1461.